ARPA Image Understanding
Workshop 1996

Obstacle Avoidance Via Depth From Focus *

Illah R. Nourbakhsh, David Andre, Carlo Tomasi, and Michael R. Genesereth
Computer Science Department, Stanford University
Stanford, CA 94305

Abstract

A critical challenge in the creation of autonomous mobile robots is the reliable detection of moving and static obstacles. In this paper, we present a passive vision system that recovers coarse depth information reliably and efficiently. This system is based on the concept of depth from focus, and robustly locates static and moving obstacles as well as stairs and dropoffs with adequate accuracy for navigation. We describe an implementation of this vision system on a mobile robot as well as real-world experiments both indoors and outdoors. These experiments have involved several hours of continuous and fully autonomous operation in crowded, natural settings.

1 Introduction

A mobile robot must be able to avoid both static and moving obstacles in its path. This essential task often relies on sonars to provide distance measurements in the immediate vicinity of the robot. Sonars are inexpensive, relatively reliable, and require little computation to process the data they provide. However, they are also limited in range and resolution, and are sensitive to false echoes caused by specular reflections. In some applications, sonars may also be undesirable because of interference problems with other sonar systems and because they are active devices that send ultrasounds into the environment. Infrared devices share many of these problems with sonars.

Vision systems are passive, have in principle no resolution limitations and have potentially a very long depth range. However, vision systems for the computation of depth from stereo correspondence or image motion are still in their infancy. They make sometimes unrealistic assumptions about the environment and even when the latter are satisfied these systems are brittle and sensitive to image measurement errors.

In this paper, we show that depth from focus provides an inexpensive and reliable sensing method for obstacle avoidance. The idea of depth from focus is not new and is even incorporated in some cameras in the consumer market. However, our particular embodiment of this concept is the first to make depth from focus useful for robot navigation, relatively inexpensive, and remarkably reliable. A $5,000 prototype of our system has so far accumulated more than twenty hours of navigation with no failure in demanding indoor and outdoor environments under varying lighting conditions, crowds of people walking past and towards the robot, and treacherous steps and obstacles all around. Our robot gracefully coasts around tables and chairs, mingles with people who pay little or no attention to it, and happily spins around when children hold hands around it singing "ring around the roses."

Our system performs no convolutions except those computed for free by defocused lenses. It has no explicit mathematical model of how defocusing alters an image, and the processing is simple enough to be carried out at frame rate on a personal computer. Three images taken from the same viewpoint but with lenses focused at different distances are compared for sharpness. The setting of the corresponding camera's focal length is taken to be the desired depth at any one image position. Although depth resolution could be increased by using more cameras, the simple strategy used to control our robot only needs to tell close from medium-distance from far at a relatively coarse grid of image regions. Any control strategy must eventually compress depth information into a few bits of information, since the choices for the robot's action are limited. Consequently, determining how much

*This research is supported through ARPA grant DAAH01-95-C-R009, and by NSF grants IRI-9509149 and IRI-9506064.

depth information is required before building the depth sensing module guarantees that all and only the necessary information is computed.

In summary, this paper is about an extremely simple idea. Simplicity itself is the point, as it yields at the same time efficiency and reliability. Our experiments, summarized in this paper and recorded on long video tapes, show our depth-from-focus system to be an alternative to sonars for its passive nature, greater accuracy, longer distance range, high reliability, and low computational cost. This is one of the first examples of a vision algorithm that is hard to defeat.

In section 2, we introduce the general idea of depth from focus and survey some of the previous work in this area. Section 3 describes the details of the implementation, and section 4 discusses our experiments. We conclude in section 5 with some general remarks.

2 Depth From Focus

The focusing ring of a modern autofocus camera provides approximate depth information about the object in the center of the camera's field of view. One could walk around with such a camera and avoid obstacles using the position of the focusing ring as a range sensor.

The robustness and simplicity of active autofocus explain the commercial success of this "active depth sensor." Unfortunately, autofocus technology has significant limitations when applied to mobile robotics. The focusing ring moves slowly and, most importantly, the autofocus system yields only one depth value for the entire field of view. In contrast, our goal is to recover depth across the entire image while taking advantage of the intuition behind many autofocus systems.

Determining depth from focus requires measuring the amount of defocus throughout the image. Computing defocus is hard because objects do not have the same inherent degree of sharpness. Therefore, an edge that appears blurred can be the result of either defocus or a soft-edged object.

A limitation that depth from focus shares with almost all other passive vision systems, including stereopsis and shape-from-motion, is that the scene must have texture or edges. Happily, natural and artificial objects are replete with texture. But depth from focus has an important advantage over stereo and motion: there is no correspondence problem.

In addition, all passive vision systems have an array of advantages over active ranging systems such as laser rangefinders, sonars, and active infrareds. Passive systems have no intrusive component, no interference problems, and no physical anomalies that accompany active ranging such as a sonar signal's specular reflection and infrared's reflectivity eccentricities based on object color and texture.

The problem of measuring defocus has been the core challenge of the depth from focus community. Early research made simplifying assumptions to work around this persistent problem. For instance, [Grossman, 1987] assumes that all objects have sharp edges.

A more practical solution is to shine illumination patterns on the scene, measuring the defocus of the patterns which have a known sharpness [Rioux and Blais, 1986]. The recent work of [Nayar *et al.*, 1995] has improved on this active approach, resulting in depth map recovery with extremely high precision and at speeds of 30 Hz. Nayar has optimized the illumination pattern and has minimized registration error by emitting the illumination pattern along the same optical path as the incoming image. As with all active illumination methods, this solution is of limited applicability in natural, outdoor environments where the emitted radiation can be either harmful or washed out by solar radiation.

Other recent work has focused on recovering depth by measuring defocus using the relative blurring between two images of the same scene. Pentland introduced the concept of performing inverse filtering in the spatial frequency domain to recover the local defocus operator [Pentland *et al.*, 1989]. In his work, two images of the same scene are taken, one with a pinhole aperture and one with a large-diameter aperture for shallow depth of field. He measures the change in defocus between two corresponding areas in the two images and thus computes the distance.

Pentland has achieved very good results, citing speeds of up to 8 frames per second. Others have improved on accuracy by using more exact defocus models based on diffraction optics [Bove, 1989]. These methods do suffer from several drawbacks. They require significant computational resources to achieve real-time performance because of the need to perform convolutions and filtering on the images.

Furthermore, the methods have generally been tested in constrained, static environments and over fairly shallow ranges of depth. For example, [Ens and Lawrence, 1991] cites results for depth map recovery over a 15 centimeter range. [Darrell and Wohn, 1988] provides experimental data that appears to encompass a comparable range although the specific distances are not disclosed. In fact, Pentland's scene, which is 1 meter cubed, appears to define the outer size limit of this body of work.

In contrast to the above work, [Krotkov, 1987] initiated the depth from focus approach, in which a large number of images with different focus operators is used to estimate the maximum focus point. Krotkov's approach also requires a static scene because the method filters intensities based on temporal averaging. Furthermore, depth is recovered for only one window in the image, whereas a mobile robot requires a multivalued depth map in order to navigate around obstacles smoothly.

Pyramid-based Depth from Focus [Darrell and Wohn, 1988] does create a depth map based on a large number of images. The authors acquire the images using one lens with a servo-controlled focusing ring. By acquiring between 8 and 30 images and interpolating the sharpness of objects over distance, they achieve excellent accuracy. Like its predecessors, this method involves nontrivial computational resources and requires a static scene.

A recent, successful vision system reported by [Krotkov and Bajcsy, 1993] combines focus and stereo ranging to achieve reliable depth over a range of two meters. Insofar as Krotkov and Bajcsy demonstrate that a cooperative ranging system is more reliable than the sum of its contributing technologies, our very simple depth from focus system can be coupled cheaply and effectively with stereo at the acquisition level and with other ranging technologies found on robots such as sonar, infrared and laser-rangefinder to create a more reliable whole.

Current research has produced depth from focus (and defocus) systems capable of high degrees of accuracy at a high computational cost. Our research takes a step away from this attitude by abandoning inverse filtering altogether to decrease computational cost dramatically while also exchanging precision in favor of simplicity and robustness. Our method does not challenge the above research results in their areas of expertise; rather, it achieves very successful results in the domain of mobile robotics, an area that the depth from focus community has not addressed.

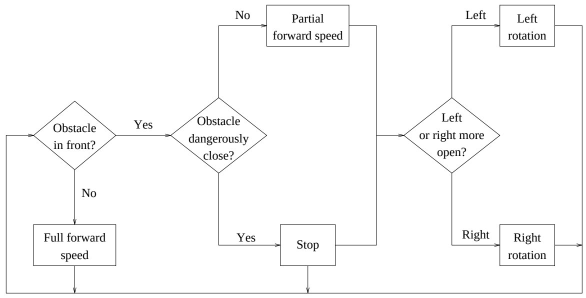


Figure 1: Robot control flowchart.

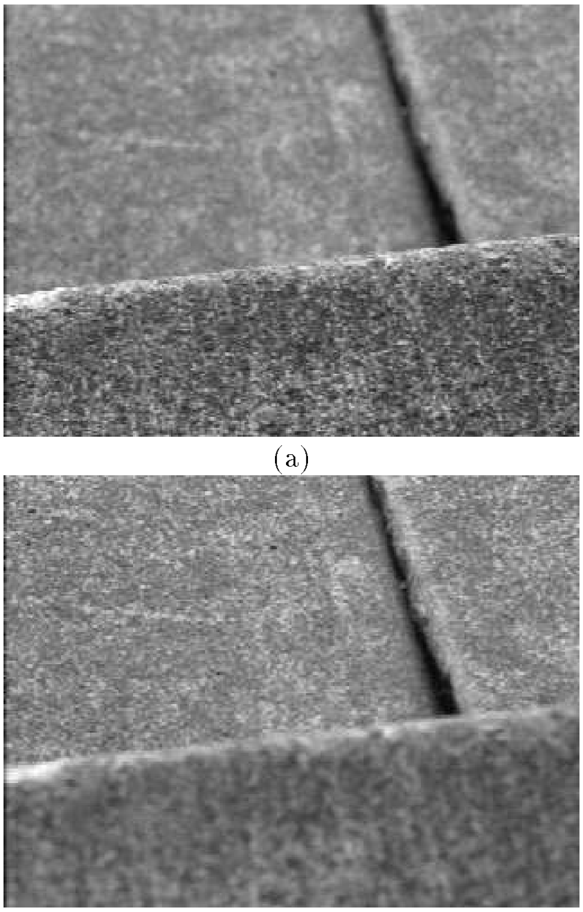
3 Categorization

In building a simple system for robot navigation, it is important to investigate the minimum requirements of a perceptual system – what is the simplest sensing system that will allow robust obstacle avoidance behavior? Figure 1 suggests a flowchart for a simple robot control strategy. Although simpler strategies may be possible, the one shown allows for some degree of smooth behavior, as it slows down and begins to turn away from objects rather than coming to jarring halts in front of them.

What granularity is required of a vision system for this strategy? At a minimum, the sensing system has to differentiate between left and right and between close, medium, and far. That is, there are only three categories of distance that are crucial to this control strategy. Of course, the distances corresponding to each distance category depend upon scene dynamics and the robot's maximum forward speed. It makes sense to ensure that the distance value corresponding to "close" be sufficiently large to allow the robot to come to a complete stop from its intended forward speed, given the cycle times of the robot control system and vision system.

Our depth from focus method capitalizes on the simplicity of this control strategy by categorizing a scene into only three levels of depth. The resulting algorithm is much simpler than many previous depth from focus algorithms [Pentland *et al.*, 1989; Ens and Lawrence, 1991]. In fact, we depend upon the convolution that is performed instantly and for free by the defocusing lens.

We begin by simultaneously recording n images of the same scene using n cameras. Ideally, the images would be identical except for the position of the focusing rings during image capture. In practice, this would require a light splitter to allow all cameras the same scene perspective. Instead of using such a splitter, we grouped the



(a)

(b)

Figure 2: Pictures of a concrete step at two different focus settings.

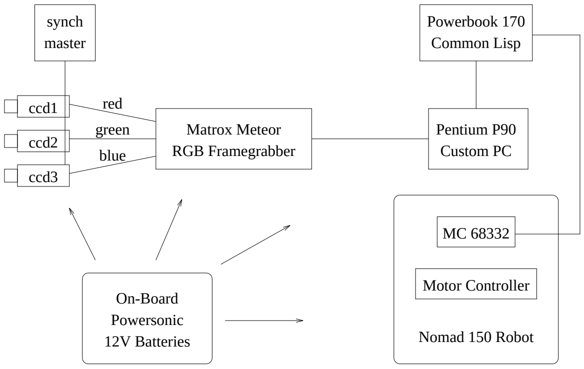


Figure 3: Architecture of the mobile robot.

cameras closely and introduced a small vergence to minimize the image shift error. Although the image shift is still quite evident, this approach has proven to be successful in our experiments.

The scene is divided into regions, and the best distance for each region can be computed by determining the image that provides the best focus. This results in a depth map of the scene, with a depth granularity equal to the number of images and width/height granularity based on the number and shape of regions.

Figure 2 shows two focus points of a concrete sidewalk step. The closer portion of the step, which occupies the lower half of the images, is best in focus in image 2 (a), indicating that the focusing ring position for this image (40 inches) is a better distance estimate to this step than the focus position of image 2 (b) (55 inches).

We compute the sharpness of a region as the sum of the absolute values of the intensity differences between all horizontally neighboring pixels in the region. In order to compute this sharpness measure, the algorithm must make only one pass over the pixel values. The entire algorithm is linear in cost with the number of pixels and the number of regions. Thus, the categorization method we propose yields a computationally inexpensive method for obstacle avoidance.

4 Implementation

A schematic diagram of the mobile robot system is shown in Figure 3. This obstacle-avoiding robot consists of three subsystems: motor, vision, and motion control. The motor subsystem, which also serves as a chassis for the entire robot, consists of a Nomad 150 robot (Nomadic Technologies, Inc.). This cylindrical robot has three degrees of freedom: it can translate forward and backward; it can rotate its wheel base; and it can rotate its turret (the upper third of the robot, which moves separately from the wheel base).

The task of the vision and control system is to provide a desired velocity to the robot for each of the degrees of freedom, approximately once every 0.25 seconds. The robot's sonar ring and position encoders were disabled to create a purely vision-driven robot.

The vision subsystem is entirely aboard the Nomad 150. It consists of four Sony XC77 2/3" CCD cameras (768 x 493), a neutral density filter, three lenses, three junction boxes, a Matrox Meteor RGB framegrabber, and a custom Pentium computer with the Triton chipset and fastPCI toolkit. The three lenses are identical 12 mm lenses with a field stop range of 0.3m to infinity. The entire system is powered by onboard 12-volt PowerSonic batteries and can run for several hours. The total cost of this vision hardware is approximately $5,500.

Three cameras are angled down from the horizontal at approximately 35 degrees, with a small vergence to minimize the image shift error. A one-to-three image splitter would have been an even better optical solution to the problem of minimizing perspective shift between the three images, although such an optical solution proved unnecessary. The focusing rings are set to positions of 0.4m, 1m, and 2m, corresponding to close, medium, and far categories. The three CCD cameras are synchronized by being slaved to a fourth Sony CCD's synch output. The three signals are fed directly into the separate R, G, and B inputs of the color framegrabber. The framegrabber digitizes the signals simultaneously and then stores the "color" image on the Pentium computer's main memory using DMA over the PCI bus, which is capable of a transfer rate greater than 100 megabytes per second. This high transfer speed enables us to recover 8 bits of intensity for each "color," or image. Our actual frame rate is approximately 15 frames per second. However, because of memory access limitations under DOS, we were able to capture and access the image data at only 4 frames per second.

In our experiments, we found that breaking the 640 by 480 images into an 8x5 depth map provided for reasonable performance and sufficient granularity for our task. The depth map contains 40 regions, each assigned 2, 1, or 0, corresponding to close, medium, and far.

As a result of the camera's downward pitch, the "expected" obstacle-free depth map contains two bottom rows of 'medium' and three top rows of 'far' categorization. If the depth map contains more than four 'far' values in the bottom two rows, there is strong evidence that there is an object farther away than the floor very near the robot (i.e. a dropoff of some sort). The robot stops in this circumstance, turns 180 degrees, and begins moving again. Using the bottom two rows rather than just the bottom row enables the robot to stop earlier as it approaches the step because the dropoff is first detected higher on the image.

In the absence of a step, the algorithm first sets any 'medium' (1) values in the bottom two rows to 'far' (0), and then sums the values in each of the eight columns. If any of these sums is larger than 1, then there must have been at least two 'medium' values or one 'close' value in that column, and the control system turns away from impending doom. The robot chooses the direction by comparing the sum of the left four columns with the sum of the right four columns.

The robot's rotational velocity in degrees per second is 5 times the sum of all depth map columns. The translational velocity in inches per second is $12 - (m \times 35)$, where m is the maximum of the depth map column sums. The rotational velocity is governed to remain in the range -30 to 30 while the translational velocity is governed to remain between 0 and 25.

5 Experiments

Initial mobile tests were conducted in several lecture rooms at the Computer Science Department. These rooms have bright, diffuse lighting and a variety of obstacles: chairs, tables, and people. The robot detected all obstacles in this setting and was able to move at a speed of 10 inches per second reliably. We were surprised to discover that the robot, which is 19 inches wide, navigates easily through standard 33 inch doorways– despite having only 6.5 inches clearance on each side.

Further indoor tests were conducted in the hallways and lounge areas of the Computer Science Department. These test were the most challenging: lighting conditions and wall texture vary greatly throughout the area. Additional risks included two open staircases and slow-moving students who actively tried to confuse the robot into falling down the stairs.

The robot performed extremely well in this complex indoor domain, avoiding the staircase as well as the students. The robot can reliably navigate from inside a classroom, through the doorway, into the hallway, past the stairs, and into the lounge with perfect collision avoidance in spite of moving students. We executed this ten minute sequence, then allowed the robot to wander the lounge for an additional ten minutes several times. In all three runs, the robot operated fully autonomously and the only environmental modification involved the removal of one coffee table in the lounge that violates our 'beheading' constraint. Average speeds in this domain were approximately 8 inches per second.

The transition to outdoor test domains introduced novel environmental characteristics. The outdoor world contains extremely intense and direct lighting, forcing us to place sunglasses (neutral density filter gel) on the robot's 'eyes' to preserve the shallow depth of field associated with a wide iris aperture.

The outdoor environment also contains an abundance of single steps which have only 7 inch drops. For safety, the visual system would have to detect all such steps without error. Furthermore, the floor of the outdoor arcade, where we ran our first outdoor tests, is composed of twelve inch tiles with discrete edges and a checkerboard coloring pattern rather than the homogeneous texture of indoor carpeting.

Testing at these outdoor steps proved that the step detection feature is extremely reliable and does not make assumptions about the homogeneity of floor texture. The robot was able to detect the steps and stop safely in all cases, even during oblique approach angles of up to 60 degrees. Over several weeks of testing, accumulating more than 15 hours of outdoor time, the robot detected dropoffs and static obstacles with 100% reliability. Furthermore, false positive detection of steps proved to be essentially nonexistent.

Our final experimental result involved an outdoor demonstration of the robot for members of the Computer Science department and researchers from industry. On June 12, 1995, the robot was placed in Memorial Court. This concrete-floored “playground” is bordered by a dropoff along one edge, stairs leading up along the other edge, and bushes and pillars along the other two edges.

We invited participants of all ages to interact with the robot, stepping in its way and controlling its path by “herding” it. They were also instructed to herd the robot toward the dropoff to test its reliability there. During a continuous two-hour demonstration, the robot interacted completely autonomously with twenty to forty participants at a time. Children even held hands, successfully encircling the robot to play 'ring around the roses.' The robot approached the dropoff and the staircase more than fifteen times, detecting them with 100% accuracy.

The robot's sensory and effectory loops were sufficiently fast that participants, including small children, were able to easily herd the robot from place to place simply by walking alongside it. The robot moved at a speed of 10 inches per second, which is approximately slow walking pace. Over the course of the demonstration, the robot came in contact with no static obstacles and contacted a moving obstacle (i.e. a human) only once.

6 Conclusion

Experimental results demonstrate that this vision system enables robust obstacle avoidance in a wide variety of environments. Significantly, the implementation does so with relatively inexpensive equipment and a highly granular but reliable depth map. Future work will extend the current set of experiments by implementing the Depth Categorization module on a robot with pan and tilt degrees of freedom. This will allow removal of the 'beheading' constraint and enable the construction of larger depth maps as well as the implementation of directed attention approaches.

References

[Bove, 1989] V. M. Bove. Discrete fourier transform based depth-from-focus. Image understanding and machine vision, 1989. technical digest series v. 14, Optical Society of America and Air Force Office of Scientific Research, June 1989.

[Darrell and Wohn, 1988] T. Darrell and K. Wohn. Pyramid based depth from focus. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 504–509, 1988.

[Ens and Lawrence, 1991] J. Ens and P. Lawrence. A matrix based method for determining depth from focus. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1991.

[Grossman, 1987] P. Grossman. Depth from focus. *Pattern Recognition Letters*, 5:63–69, January 1987.

[Krotkov and Bajcsy, 1993] E. Krotkov and R. Bajcsy. Active vision for reliable ranging: Cooperating focus, stereo, and vergence. *International Journal of Computer Vision*, 11(2):187–203, 1993.

[Krotkov, 1987] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1:223–237, 1987.

[Nayar *et al.*, 1995] S. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. In *IEEE International Conference on Computer Vision*, pages 824–831, 1995.

[Pentland *et al.*, 1989] A. Pentland, T. Darrell, M. Turk, and W. Huang. A simple, real-time range camera. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 256–261, 1989.

[Rioux and Blais, 1986] M. Rioux and F. Blais. Compact three-dimensional camera for robotic applications. *Journal of the Optical Society of America*, 3(9):1518–1521, September 1986.